Bayesian integration of visual and auditory signals for spatial localization

Peter W. Battaglia, Robert A. Jacobs, and Richard N. Aslin

Department of Brain and Cognitive Sciences and The Center for Visual Science, University of Rochester, Rochester, New York 14627

Received September 11, 2002; revised manuscript received January 21, 2003; accepted February 20, 2003

Human observers localize events in the world by using sensory signals from multiple modalities. We evaluated two theories of spatial localization that predict how visual and auditory information are weighted when these signals specify different locations in space. According to one theory (visual capture), the signal that is typically most reliable dominates in a winner-take-all competition, whereas the other theory (maximum-likelihood estimation) proposes that perceptual judgments are based on a weighted average of the sensory signals in proportion to each signal's relative reliability. Our results indicate that both theories are partially correct, in that relative signal reliability significantly altered judgments of spatial location, but these judgments were also characterized by an overall bias to rely on visual over auditory information. These results have important implications for the development of cue integration and for neural plasticity in the adult brain that enables humans to optimally integrate multimodal information. © 2003 Optical Society of America

OCIS codes: 330.0330, 330.1400, 330.4060, 330.7320.

1. INTRODUCTION

The ability to localize a stimulus in the environment is based on a complex mapping of sensory signals that leads to a perceptual judgment. For example, auditory localization relies, in part, on binaural time-of-arrival differences that must be scaled by the distance between the two ears to provide a "correct" interpretation of the stimulus location with respect to the head. Similarly, visual localization relies, in part, on the coordinates of retinal stimulation, the position of the eye in the orbit, and the orientation of the head on the body. Thus stimulus localization entails the integration of multiple sources of sensory and motor information.[1]

In most circumstances, events in the environment provide consistent cues to spatial location; a mouse running across a field is visually, auditorily, and tactually located in the same position in space when a barn owl swoops down to capture it. However, not all events in the environment are characterized by consistent cues.[2] For example, in a movie theater the visual information is located on the screen whereas the auditory information often comes from loudspeakers located to the side of the screen. Nevertheless, we perceive the sound to originate from the location of the visual stimulus (e.g., the moving lips of a face or the crash of an automobile). This is an example of "visual capture" in which the visual information for spatial location dominates completely the conflicting auditory information.[3,4] Knudsen and his colleagues have shown in the barn owl that vision dominates audition when these two sources of information are artificially put into conflict.[5,6] Juvenile barn owls whose auditory cues to the location of a sound are altered (with a monaural earplug) or whose visual cues to object location are altered (with displacing prisms) recalibrate the relationship between sight and sound, with vision dominating audition.

Two models have been proposed to account for how observers make perceptual judgments when signals from different modalities are in conflict. One model proposes that the signal that is typically most reliable dominates in a winner-take-all competition, and an observer's judgment is based exclusively on that dominant signal. In the context of spatial localization based on visual and auditory signals, this model is called visual capture because localization judgments are dominated by visual information. The other model proposes that perceptual judgments are based on a blend of information arising from multiple modalities.[7] Several investigators have recently examined whether human adults combine information from multiple sensory sources in a statistically optimal manner.[8–11] With certain mathematical assumptions, an optimal model of sensory integration has been derived based on maximum-likelihood estimation (MLE) theory. Specifically, the model assumes that the sensory signals are statistically independent given a value of a property of a scene and that an observer's estimate of the value of the scene property given a sensory signal has a normal distribution. In the engineering literature, the MLE model is also known as a Kalman filter.[12,13] According to this model, a sensory source is reliable if the distribution of inferences based on that source has a relatively small variance; otherwise the source is regarded as unreliable. More-reliable sources are assigned a larger weight in a linear-cue-combination rule, and less reliable sources are assigned a smaller weight. Thus visual capture is simply a special case in which one highly reliable cue (vision) is assigned a weight of one and a less reliable cue (audition) is assigned a weight of zero.

Figure 1 illustrates two hypothetical situations in which visual and auditory signals provide different information about the location of an event. In the two graphs in this figure, the horizontal axis indicates a spatial loca-

1084-7529/2003/071391-07$15.00

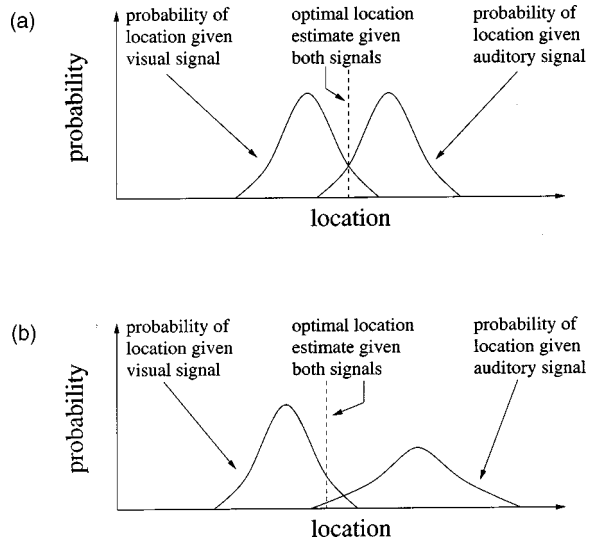

Fig. 1. Optimal model of sensory integration based on MLE theory. (a) Visual and auditory signals are equally reliable indicators of event location. (b) Visual signal is a more reliable indicator of event location.

tion and the vertical axis plots the probability that the event occurred at a location based on one or more sensory signals. The leftmost normal distribution in each graph gives the probability distribution of locations based on the visual signal; the rightmost probability distribution is for locations based on the auditory signal.

In Fig. 1(a), the two distributions have equal variances, indicating that visual and auditory signals are equally reliable information sources about location. In this case, the statistically optimal way of integrating visual and auditory signals is to linearly average the best location estimate based on the visual signal with the best location estimate based on the auditory signal, where the two location estimates (e.g., the peaks of the distributions) are given equal weight. The dashed line in the graph shows the optimal estimate of location based on both sensory signals.

In Fig. 1(b), the probability distribution based on the visual signal has a smaller variance than the distribution based on the auditory signal. In this case, the visual signal is regarded as more reliable. In computing the optimal estimate of location based on both signals, a linear average of the best location estimates based on the individual signals assigns a larger weight to the estimate derived from visual information.

Mathematically, the MLE model of sensory integration is characterized in the following way. Let L denote a possible location of an event, and let v and a denote the values of the visual and auditory signals. In addition, let L_v^* denote the best location estimate based on the visual signal [this is the location L that maximizes the probability of a location value given the visual signal, $P(L|v)$], let L_a^* denote the best location estimate based on the auditory signal [the location L that maximizes $P(L|a)$], and let L^* denote the optimal location estimate based on both visual and auditory signals [the location L that maximizes $P(L|v,a)$]. Then it can be shown that the optimal location estimate based on both signals can be computed as follows:

$$L^* = w_v L_v^* + w_a L_a^*, \tag{1}$$

where

$$w_v = \frac{1/\sigma_v^2}{1/\sigma_v^2 + 1/\sigma_a^2} \text{ and } w_a = \frac{1/\sigma_a^2}{1/\sigma_v^2 + 1/\sigma_a^2} \tag{2}$$

and σ_v^2 and σ_a^2 are the variances of the distributions $P(L|v)$ and $P(L|a)$, respectively.[14]

Do human observers localize events on the basis of visual and auditory signals in a way that is best predicted by the visual capture model or by the MLE model? We conducted an experiment designed to evaluate this question. Surprisingly, the results indicate that both models are partially correct and that a hybrid model may provide the best account of subjects' performances. We examined the extent to which subjects use visual and auditory information to estimate location when the visual signal is corrupted by noise of varying amounts. As greater amounts of noise were added to the visual signal, subjects tended to use auditory information more and more. Although this trend is predicted by the MLE model, this model does not correctly predict subjects' responses. Instead, the model makes a systematic error by consistently underestimating the degree to which subjects made use of visual information. That is, subjects seem to be biased to use visual information to a greater extent than predicted by the MLE model, a bias that is broadly consistent with the visual capture model. Our findings are interesting because they provide a new way of thinking about a modified MLE model that is biased toward the use of visual information or, alternatively, a modified visual capture model that is made probabilistic in the manner of the MLE model. Overall, our results can be accounted for by a Bayesian model that modifies the MLE model through the addition of a prior probability distribution that leads the model to make greater use of visual information.

2. METHODS

A. Subjects

Ten subjects participated in the study. All had normal or corrected-to-normal vision and normal hearing. Subjects were naïve to the purposes of the study. All subjects gave informed consent according to procedures approved by the University of Rochester Research Subjects Review Board.

B. Stimuli and Experimental Apparatus

The auditory signal used in the experiment was a broadband noise burst filtered to eliminate onset and offset transients and to mimic the spectral characteristics of a sound source external to the listener (with use of head-related transfer functions supplied by F. Wightman[15]). In particular, these characteristics were manipulated so that the noise appeared to originate from one of seven locations arranged along a horizontal axis spanning the width of the experimental workspace. Locations were spaced at intervals of 1.5° of visual angle. These locations are referred to as comparison locations. The stimulus was created with a Tucker–Davis Technologies RP2 signal processor and the Visual Design Studio software. The processor was connected to a Tucker–Davis Technolo-

gies HB7 headphone driver that delivered the stimulus through Sennheiser HD-265 headphones.

The visual stimulus was a random-dot stereogram of a bump, shaped like a normal distribution, protruding from a frontal parallel background surface. The bump was centered at one of the seven comparison locations. The height at the top of the bump was 150 pixels from the background (20.8 cm assuming a viewing distance of one m). The dots of the stereogram subtended 13.5 arc min, and the density was ~6 dots per square degree. The visual signal was corrupted by one of five levels of noise. The noise was created by scattering a percentage of dots at random depths in the workspace instead of being placed on the background or the bump.[10,16] These noise dots were placed at random depths in the interval from 15 pixels behind the background to 15 pixels in front of the peak of the bump (where 15 pixels is 2.08 cm assuming a viewing distance of one m). The five noise levels randomized 10%, 23%, 36%, 49%, and 62% of the dots. The visual stimulus was displayed on two monitors mounted in a Virtual Research V6 head-mounted display system. Each monitor had a resolution of 640 × 480 pixels and a refresh rate of 60 Hz. The effective viewing distance was approximately 1 m.

C. Procedure

The experiment was conducted in two phases. Trials in the first phase used signals from a single sensory modality, whereas trials in the second phase used signals from both modalities. The goal of the first phase was to measure subjects' localizations when exposed to either an auditory signal or to a visual signal corrupted with one of five levels of noise. On the basis of these trials, the parameter values of the MLE model were estimated for each subject at each visual noise level (see Appendix A for mathematical details). The MLE model used the data from single-modality trials so that predictions could be made about subjects' judgments when exposed to auditory and visual signals simultaneously. The accuracy of these predictions was evaluated in the second phase of the experiment when subjects localized events on the basis of both signals, which were presented in different spatial locations.

In each single-modality trial, subjects observed stimuli in two temporal intervals and judged whether the event depicted in the second interval was located to the left or the right of the event depicted in the first interval. One stimulus, called the standard, always depicted an event at the center of the experimental workspace. The other stimulus, called the comparison, depicted an event at one of the seven comparison locations. Figures 2(a) and 2(b) illustrate the auditory-only and visual-only trials. The standard stimulus is shown on the left of each figure and the comparison stimulus on the right.

In each multimodality trial, the standard stimulus included both visual and auditory signals, but these signals depicted events at different locations. The visual signal corresponded to an event at −1.5° to the left of the workspace center and the auditory signal an event at 1.5° to the right. This discrepancy was introduced so that we could measure the relative degree to which subjects relied on visual versus auditory information when localizing events. For example, a subject who localized the event in the standard stimulus at −1.5° would be basing that judgment entirely on the visual signal, a localization at 1.5° would indicate that the judgment was based entirely on the auditory stimulus, and a localization at 0° would suggest that the subject weighted visual and auditory information equally. With one possible exception, all subjects reported being unaware of the discrepancy between the visual and auditory signals in the standard stimulus. These signals were spatially coincident in the comparison stimulus [see Fig. 2(c)].

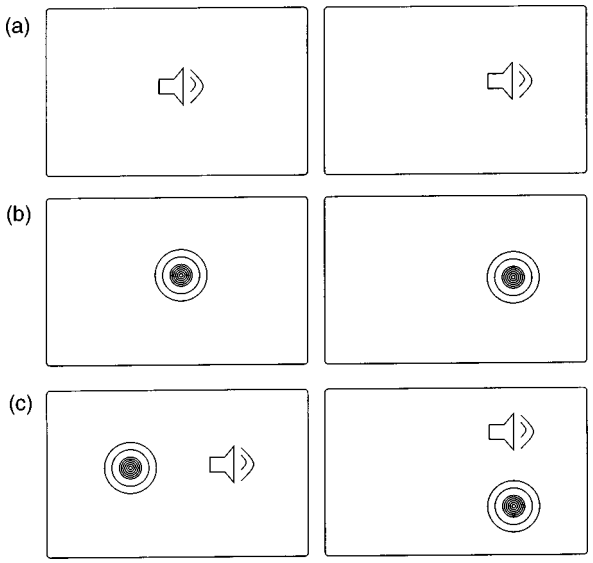

Fig. 2. Schematic illustration of single-modality and multimodality trials. The standard stimulus is shown on the left and the comparison stimulus is on the right. For simplicity, the comparison stimulus is shown only at one of the seven possible locations at which it could appear. (a) Auditory-only trial. (b) Visual-only trial. (c) Visual–auditory trial.

Subjects participated in two experimental sessions on successive days. In the first session they performed three blocks of practice trials consisting of 35 trials each of either auditory-only, visual-only, or visual-auditory trials. The data from practice trials were not used in the study. Subjects then performed 105 auditory-only trials followed by 525 visual-only trials (105 trials at each noise level × 5 noise levels; trials with different noise levels were randomly intermixed). In the second session, subjects performed 525 visual–auditory trials (again, trials with different noise levels were intermixed).

On each trial, standard and comparison stimuli were presented in random order to eliminate an anchoring bias. Subjects used key presses to indicate their judgments of whether the event depicted in the second stimulus was to the left or the right of the event depicted in the first stimulus. Stimuli were presented for 500 ms, and there was a 500-ms delay between the presentations of the first and the second stimuli.

3. RESULTS

The results for one subject on the auditory-only trials are shown in Fig. 3. The horizontal axis shows the comparison locations (in degrees of visual angle away from the

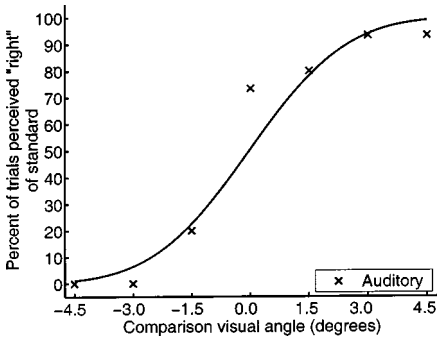

Fig. 3. Results for one subject on the auditory-only trials. The horizontal axis shows the comparison locations (in degrees of visual angle away from the center of the workspace), and the vertical axis shows the percentage of trials in which the subject judged the comparison stimulus as depicting an event located to the right of the event depicted in the standard stimulus. The curve fitted to the data points is a cumulative normal distribution.

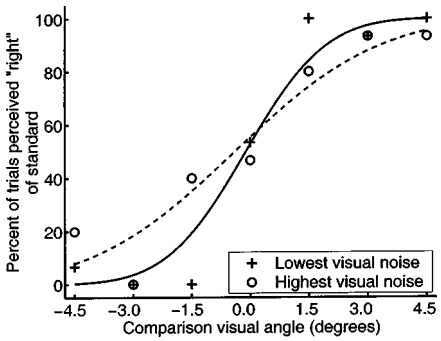

Fig. 4. Results for one subject on the visual-only trials. The solid and dashed curves are cumulative normal distributions fitted to the data points in the lowest-noise and highest-noise conditions, respectively.

center of the workspace), and the vertical axis shows the percentage of trials in which the subject judged the comparison stimulus as depicting an event located to the right of the event depicted in the standard stimulus. A cumulative normal distribution was fitted to the data points, and the mean and variance of this distribution were used by the MLE model as described in Eqs. (1) and (2).

The results for the same subject on the visual-only trials are shown in Fig. 4. The two sets of data points, and their corresponding best-fitting cumulative normal distributions, are for the trials when the visual signal was corrupted by the lowest and the highest amounts of noise. Because the distribution fitted to the data points in the highest-noise condition (dashed curve) has a larger variance than the distribution for the lowest-noise condition (solid curve), we can conclude that greater amounts of noise in the visual signal increased the uncertainty in the subject's localization judgments.

On the basis of the data in Figs. 3 and 4, the MLE model can be used to predict the subject's responses on trials that contain conflicting visual and auditory signals. The subject's responses had a comparatively small variance when the visual signal was corrupted by a small level of noise, as measured by the cumulative normal distribution fitted to the subject's data. Therefore the model predicts that the subject should weight visual information highly when both signals are available and the visual signal has little noise [i.e., w_v is assigned a large number in Eqs. (1) and (2)]. When the visual signal is highly corrupted, however, the subject's responses had a larger variance. Consequently, the model predicts that the subject should weight visual information to a smaller degree and auditory information to a greater degree (i.e., w_v is assigned a smaller number and w_a a larger number) when both signals are available and the visual signal is noisy. These predictions were tested in the multimodality trials.

Figure 5 shows the results on the visual–auditory trials for the same subject as discussed above. The stars and the solid curve are for the case when the visual signal was corrupted by the lowest level of noise, and the squares and the dashed curve are for the highest-noise condition. The dependent measure of spatial localization is the mean of each cumulative normal distribution (the point where the distribution crosses 50%), which is commonly referred to as the point of subjective equality (PSE). In the lowest-noise condition, the subject's PSE is approximately $-1.1°$, which is very close to $-1.5°$, the event location depicted by the visual signal in the standard stimulus. Consequently, we can conclude that this subject's localizations were strongly dominated by information from the visual signal when both visual and auditory signals were available and the visual signal had only a small level of noise. In contrast, the subject showed a different pattern in the high-noise condition. In this case, the PSE is approximately 0.1°, which is almost exactly in the middle of the locations depicted by the visual and auditory signals of the standard stimulus. We can conclude, therefore, that this subject's localizations were based on visual and auditory information in equal amounts when the visual signal was corrupted by a large level of noise. The shift in the subject's dominant reliance on visual information in the lowest-noise condition to a balanced reliance on both visual and auditory information in the highest-noise condition is in qualitative agreement with the predictions of the MLE model.

Although there was variability in the performances of different subjects, the subject discussed above is typical. Figure 6 shows the average PSE over all ten subjects on the multimodality trials. The horizontal axis represents the visual noise level (1, lowest level; 5, highest level),

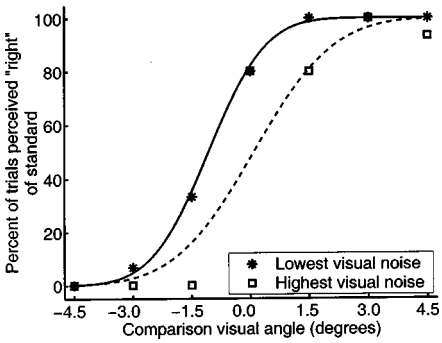

Fig. 5. Results for one subject on the visual–auditory trials. The solid and dashed curves are cumulative normal distributions fitted to the data points in the lowest-noise and highest-noise conditions, respectively.

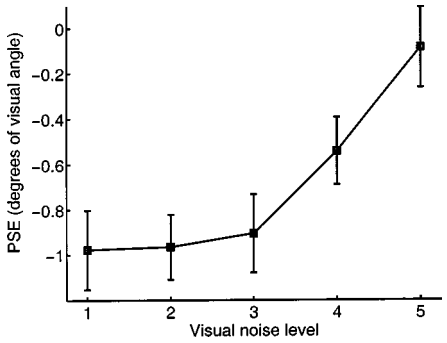

Fig. 6. Average PSE over all ten subjects on the visual–auditory trials. The horizontal axis represents the visual noise level (1, lowest level; 5, highest level), and the vertical axis gives the average PSE in degrees of visual angle (the error bars give the standard errors of the means).

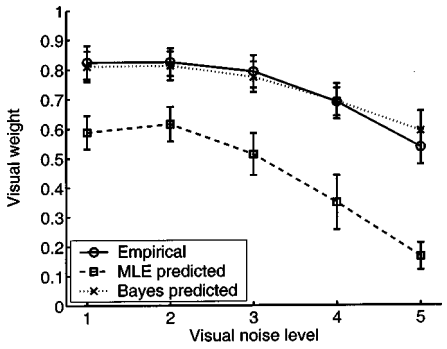

Fig. 7. Average visual weights over all ten subjects on the visual–auditory trials. The horizontal axis represents the visual noise level (1, lowest level; 5, highest level), and the vertical axis gives the average visual weight (the error bars give the standard errors of the means).

and the vertical axis shows the average PSE in degrees of visual angle (the error bars give the standard errors of the means). On average, PSEs were close to $-1.5°$ at the lowest visual-noise level, indicating that subject's judgments were based mostly on the visual signal. At the highest noise level, PSEs were close to 0°, indicating that subjects used visual and auditory information in roughly equal amounts in this condition.

If we assume that subjects use a linear-cue-combination rule for integrating visual and auditory information [Eq. (1)], then we can estimate the degree to which each subject used visual and auditory information (the values of the visual and auditory weights w_v and w_a) on the basis of their responses on the multimodality trials (see Appendix A for further details). Figure 7 shows the average visual weights across all ten subjects (the auditory weight is one minus the visual weight). The horizontal axis represents the visual noise level, and the vertical axis shows the visual weight. The estimated weights based on subjects' empirical data are given by the open circles connected with a solid line. At the lowest visual noise level, visual weights were large (approximately 0.8). As the amount of noise in the visual signal increased, visual weights decreased monotonically. The open squares connected with a dashed line give the visual weights predicted by the MLE model. Although the general shape of the MLE model's predictions fits the data well, this model does not predict subject's responses correctly. Instead, the model makes a systematic error by underestimating the extent to which subjects made use of visual information. This bias toward the use of visual information is broadly consistent with the visual capture model and suggests that a hybrid approach that combines properties of the MLE model and the visual capture model may provide a good account of the data.

To evaluate this hybrid approach, we developed a Bayesian model that is identical to the MLE model except that it includes a prior probability distribution that leads the model to make greater use of visual information. In short, this prior probability distribution is used in estimating the variances of the cumulative normal distributions for the visual-only trials. The use of this prior causes the Bayesian model to estimate smaller values for the variances of the cumulative distributions than the MLE model. As a result, the Bayesian model estimates larger values for the visual weights at all noise levels (see Appendix A for further details). The $\times$s connected with a dotted line in Fig. 7 show the Bayesian model's predicted visual weights, which are in close agreement with the empirical weights. An appropriate test of the Bayesian model would evaluate whether the prior distribution, which we estimated on the basis of the data collected in our experiment, provides accurate predictions of observers' judgments in other stimulus contexts.

4. DISCUSSION

The visual capture and MLE models are commonplace in the scientific literature on sensory integration. A strength of the visual capture model is that it accounts for the finding that observers' perceptual judgments in multimodal situations often seem to be dominated by visual information. Its weaknesses include the fact that it fails to account for the probabilistic nature of observers' percepts. The MLE model has complementary strengths and weaknesses. It proposes an elegant statistical model of observers' multimodal percepts, but it fails to take into account observers' perceptual biases. We believe that the work reported here demonstrates the strengths and weaknesses of both of these models and highlights the fact that a hybrid approach may provide the best explanation for observers' perceptual judgments.

The MLE model hypothesizes that observers judge the reliability of a sensory signal as inversely proportional to the variance of the distribution of inferences based on that signal. It is not known, however, how observers estimate this variance. It is possible that a neural representation of a stimulus property in a scene may encode the uncertainty in sensory signals.[10] For example, consider an observer localizing an object in space on the basis of visual and auditory signals. The activities of neurons in the observer's visual cortex form a neural population code that represents an estimate of the object's location based on the visual signal as well as the uncertainty in this estimate. Similarly, a neural population code in the observer's auditory cortex represents a location estimate, and the uncertainty of this estimate, based on the auditory signal. If each population code were shaped like a normal distribution such that the mean and the variance

of this distribution represented the location estimate and the uncertainty in this estimate, respectively, then the nervous system could implement the MLE model in a direct manner. The product of two normal-shaped population codes, based on two sensory signals such as visual and auditory cues to location, is also a normal-shaped population code. The mean and variance of this new code represent the optimal location estimate according to MLE theory based on the mean of each signal and their respective uncertainties. Computational neuroscientists have made important progress in recent years in developing biologically realistic neural models that perform MLE using population codes.[17,18] An unsolved problem, but one that is currently being pursued, is to develop models that show perceptual biases through the use of prior probability distributions.[19,20]

Although observers are biased toward the use of visual information when localizing events in the world, they might not show this bias in other contexts. For example, Shams, Kamitani, and Shimojo[21,22] reported a case in which subjects seemed to be biased toward the use of auditory information: When a single visual flash is accompanied by multiple auditory beeps, the single flash is incorrectly perceived as multiple flashes. It seems possible that the MLE model can account for the diversity of experimental findings from use of visual and auditory stimuli, at least in part, by taking advantage of the fact that some estimates (e.g., of spatial location) may be more precise in the visual system whereas other estimates (e.g., of temporal variations) may be more precise in the auditory system.

Observers localizing events show a bias toward the use of visual information as a result of either evolutionary history or experience-dependent learning. If the latter, then it would be interesting to evaluate whether observers with different histories of visual experience show the same bias. For example, consider observers with impaired vision or observers who lost vision at an early age but then had some visual abilities restored later in life.[23] Would these individuals also show a bias toward the use of visual information? Such "experiments of nature" could reveal whether visual capture is under adaptive control or whether it is a hard-wired bias.

APPENDIX A: MATHEMATICAL PROCEDURES FOR ESTIMATING VISUAL AND AUDITORY WEIGHTS

A cumulative normal distribution was fitted to each subject's responses on the auditory-only trials and on the visual-only and visual–auditory trials at each of the five visual noise levels. The mean and variance of this distribution were found by MLE. Because responses on each trial were a binary event, the likelihood function is a Bernoulli function,

$$p(\mathcal{R}|\mu, \sigma^2) = \prod_{t=1}^{T} p_t^{r_t}(1 - p_t)^{1-r_t}, \tag{3}$$

where t denotes the trial number, r_t denotes the subject's response on trial t (1, event in comparison stimulus is to right of event in standard stimulus; 0, otherwise), $p_t = p(r_t = 1/\mu, \sigma^2)$ denotes the probability that $r_t = 1$ on trial t according to a cumulative normal distribution with mean μ and variance σ^2, and $\mathcal{R}$ denotes the entire set of responses for all trials. The values of μ and σ^2 that maximize Eq. (3) are the maximum-likelihood estimates of these parameters. Visual and auditory weights for the MLE model were computed by using maximum-likelihood estimates of the variances for the auditory-only and visual-only trials in Eq. (2).

The visual and auditory weights based on subjects' responses on the visual–auditory trials, referred to as empirical estimates in Fig. 7, were computed as follows. Let L_v^c, L_a^c, and L^c denote location estimates based on the visual signal, the auditory signal, and both signals in the comparison stimulus. Similarly, let L_v^s, L_a^s, and L^s denote the corresponding quantities for the standard stimulus. Assume a linear-sensory-integration rule,

$$L^c = w_v L_v^c + w_a L_a^c, \tag{4}$$

$$L^s = w_v L_v^s + w_a L_a^s, \tag{5}$$

where w_v and w_a are visual and auditory weights that are assumed to sum to 1. Maximum-likelihood estimates of these weights were the values that maximized the likelihood function given above [the right-hand side of Eq. (3)] with the exception that the probability p_t was given by a logistic function:

$$p_t = p(r_t = 1|w_v, w_a) = \frac{1}{1 + \exp[-(L_c - L_s)/\tau]}, \tag{6}$$

where τ is a scale parameter often referred to as a temperature or a slope.

The Bayesian model was identical to the MLE model described above with the following exception. For the visual-only trials, the cumulative normal distribution's mean and variance were estimated by using the values that maximized the product of the likelihood function given above [the right-hand side of Eq. (3)] and a prior probability distribution for these parameters. To keep things simple, we assumed that the mean and variance are statistically independent, that the prior distribution for the mean is a uniform distribution (meaning that all possible values are equally likely), and that the prior distribution for the variance is an inverse-gamma distribution (this is the conjugate prior distribution for a normal variance[24]). The parameters of the inverse-gamma distribution were set so that this distribution had most of its mass toward small values of the variance (the shape parameter was set to 46.0 and the scale parameter was set to 10^{-34}; these parameter values were used for all subjects). Consequently, the Bayesian model was biased toward smaller variance estimates for the visual-only trials and, thus, larger visual weight estimates.

ACKNOWLEDGMENTS

We thank F. Wightman for supplying the head-related transfer functions used in the delivery of our auditory stimuli. This work was supported by National Science Foundation research grant SBR98-73477 and by National Institutes of Health research grant R01-EY13149.

Corresponding author Robert Jacobs can be reached as follows: Department of Brain and Cognitive Sciences and The Center for Visual Science, University of Rochester, Rochester, New York, 14627-0268; phone, 585-275-0753; fax, 585-442-9216; e-mail, robbie @bcs.rochester.edu.

REFERENCES

1. B. E. Stein and M. A. Meredith, *The Merging of the Senses* (MIT Press, Cambridge, Mass, 1993).
2. R. Held, "Shifts in binaural localization after prolonged exposure to atypical combinations of stimuli," Am. J. Psychol. **68**, 526–548 (1955).
3. H. L. Pick, Jr., D. H. Warren, and J. C. Hay, "Sensory conflict in judgements of spatial direction," Percept. Psychophys **6**, 203–205 (1969).
4. R. B. Welch and D. H. Warren, "Immediate perceptual response to intersensory discrepancy," Psychol. Bull. **88**, 638–667 (1980).
5. E. I. Knudsen and M. S. Brainard, "Creating a unified representation of visual and auditory space in the brain," Annu. Rev. Neurosci. **18**, 19–43 (1995).
6. M. S. Brainard and E. I. Knudsen, "Sensitive periods for visual calibration of the auditory space map in the barn owl optic tectum," J. Neurosci. **18**, 3929–3942 (1998).
7. J. J. Clark and A. L. Yuille, *Data Fusion for Sensory Information Processing Systems* (Kluwer Academic, Norwell, Mass, 1990).
8. Z. Ghahramani, *Computation and Psychophysics of Sensorimotor Integration*, Ph.D. dissertation (Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, Mass., 1995).
9. R. A. Jacobs, "Optimal integration of texture and motion cues to depth," Vision Res. **39**, 3621–3629 (1999).
10. M. O. Ernst and M. S. Banks, "Humans integrate visual and haptic information in a statistically optimal fashion," Nature **415**, 429–433 (2002).
11. D. C. Knill and J. Saunders, "Humans optimally weight stereo and texture cues to estimate surface slant,"J. Vision **2**, 400 (2002).
12. R. E. Kalman and R. S. Bucy, "New results in linear filtering and prediction problems," J. Basic Eng. Ser. D **83**, 95–108 (1961).
13. G. C. Goodwin and K. S. Sin, *Adaptive Filtering Prediction and Control* (Prentice-Hall, Englewood Cliffs, N.J., 1984).
14. A. L. Yuille and H. H. Bülthoff, "Bayesian decision theory and psychophysics," in *Perception as Bayesian Inference*, D. C. Knill and W. Richards, eds. (Cambridge U. Press, Cambridge, UK, 1996), pp. 123–161.
15. F. Wightman, Department of Psychology, University of Wisconsin, Madison, Wisconsin (personal communication, 2000).
16. J. S. Tittle, V. J. Perotti, and J. F. Norman, "Integration of binocular stereopsis and structure from motion in the discrimination of noisy surfaces," J. Exp. Psychol. Hum. Percept. Perform. **23**, 1035–1049 (1997).
17. S. Deneve, P. E. Latham, and A. Pouget, "Reading population codes: a neural implementation of ideal observers," Nat. Neurosci. **2**, 740–745 (1999).
18. A. Pouget, R. S. Zemel, and P. Dayan, "Information processing with population codes," Nature Rev. Neurosci. **1**, 125–132 (2000).
19. S. Deneve, P. E. Latham, and A. Pouget, "Efficient computation and cue integration with noisy population codes," Nature Neurosci. **4**, 826–831 (2001).
20. S. Wu and S. Amari, "Neural implementation of Bayesian inference in population codes," in *Advances in Neural Information Processing Systems 14*, T. G. Dietterich, S. Becker, and Z. Ghahramani, eds. (MIT Press, Cambridge, Mass., 2002).
21. L. Shams, Y. Kamitani, and S. Shimojo, "What you see is what you hear," Nature **408**, 788 (2000).
22. S. Shimojo and L. Shams, "Sensory modalities are not separate modalities: plasticity and interactions," Curr. Opin. Neurobiol. **11**, 505–509 (2001).
23. H. S. Smallman, I. Fine, and D. I. A. MacLeod, "Pre- and post-operative characterization of visual function before and after the removal of congenital bilateral cataracts," Vision Res. **42**, 191–210 (2002).
24. A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis* (Chapman & Hall, London, 1995).